Spring 2017 SRCPAC Meeting Minutes
Friday, March 3rd, 2017

Attendees: Chris Marianetti (Chair), Alan Crosswell, Halayn Hescock, Marley Bauce, Victoria Hamilton, Rob Lane, George Garrett, Rob Cartolano, Gaspare LoDuca, Barbara Rockenbach, Mark Newton, Marc Spiegelman, Ryan Abernathey, Amy Nurnberger, Jennifer Brown, Raj Bose, Ingrid Richter, Kyle Mandli, Michael Weisner, George Garrett, Bob Mawhinney, Steve Bellovin, Eric Bloch, Jochen Weber, Ron Agmon, Harmen Bussemaker, Maneesha Aggarwal, Mike Tuts, Paul Blaer, Andrei Beloborodov

Chair Chris Marianetti opens the meeting by introducing himself as the new Chair of SRCPAC, and the first time that he has led a SRCPAC meeting, having taken over in July 2016 from the past Chair, Kathryn Johnston. The meeting attendees then introduce themselves.

Habanero Update

Chris introduces Kyle Mandli, the Chair of the Habanero Operating Committee, who explains that the OC met for the first time in January 2017. The Habanero Operating Committee meets to review the business rules, and whether to make any changes. Habanero launched with rules analogous to the Yeti business rules (nearly identical, save for necessary changes imposed by the new software). These rules will be re-reviewed in the Fall 2017 Habanero OC meeting.

Kyle then asks Rob Lane, Manager of Research Computing Services to review Habanero status. Rob shows a photo of the cluster – 222 nodes across 4 racks, with integrated power and water/cooling.

Habanero has three types of nodes – Standard, High Memory, and GPU – with precise nodes and cores and K80 GPUs detailed on the corresponding presentation slide. Habanero has a very fast network and storage attached. 32 research groups purchased 197 of the 222 nodes, with the remaining purchased through central University funds. Rob notes that GPU nodes have become increasingly popular with each HPC iteration.

The acquisition of Habanero was led through a Faculty RFP Committee composed of seven faculty participants, who met four times in early 2016. This Committee selected the storage and server vendors from 16 initial bids, thereafter CUIT arranged logistics.

Of the 32 research groups bought into Habanero, 14 were also on Yeti, meaning that 18 new groups joined. HPC has always supported educational activities, though informally via Yeti. For Habanero, SEAS and A&S split the cost of purchasing nodes exclusively for educational use. If any teachers would like to utilize these nodes for their courses, please email Rob at roblane@columbia.edu. Chris echoes this announcement by strongly encouraging researchers to take advantage of the Deans' investment.

Rob explains that soon Habanero will host one-year rental options (for $1,000) and a free-tier. Any interested parties should also contact Rob at roblane@columbia.edu. Rob notes a small but consistent group of renters (there are currently four renters on Yeti). Any member of the Columbia community may use the Free Tier – faculty and/or researchers just need to email Rob, while any other individuals must receive approval from a faculty member in order to gain access.

Most support is handled through email via hpc-support@columbia.edu, in addition to standard open office hours from 3-5pm on the first Monday of each month in the Science & Engineering Library (Northwest Corner Building). Increased demand would prompt RCS to increase the number of open office hours.

In addition to these office hours, the RCS team is always eager to speak to groups at any time regarding any topic of interest. RCS can accommodate 1-on-1 meetings, presentations to departments or schools, etc. To schedule a presentation, email hpc-support@columbia.edu.

RCS collaborates with the Libraries on a series of free workshops: Intro to Linux, Intro to Scripting, and Intro to HPC (all in consecutive weeks). The next workshop is Intro to HPC, with details provided in the corresponding presentation slide.

Rob additionally reviews historic daily usage on the Habanero cluster: the theoretical maximum is 120,000 core hours, which has been approached a few times. The episodes of low usage are primarily due to interruptions. More information can be found on the corresponding presentation slide. Rob notes that although Yeti was a very slow ramp-up, Habanero users jumped onto the system much quicker, likely as a substantially more homogenous system, Habanero can accommodate most jobs anywhere.

Marc Spiegelman asks for a distribution of job sizes on Habanero. Rob responds that Habanero is being driven by large jobs (requiring 50+ notes), while, conversely, Yeti is being driven by many small jobs.

Bob Mawhinney asks for the final network configuration – Rob responds EDR Infiniband, with FDR to storage.

Ryan Abernathey notes that storage is the “unsung hero” of Habanero because it is so fast and has significant potential. Rob responds that storage historically has not been receiving the attention it deserves, although Habanero has a much more high performance storage system than its Yeti counterpart.

Chris Marianetti asks whether we have sorted out building and facilities issues. Alan Crosswell responds that there are no more anticipated issues. Rob also notes that there are fewer surprises – we know when things are happening – although anticipates there will be more interruptions, and they will be communicated when they arise.

HPC Expansion and Annual Purchase Cycle

Chris takes the floor to explain the proposed timeline for expanding Habanero. He explains that this is a difficult thing to plan. The goal is to have the expansion round coincide with the hiring cycle so new faculty could use start-up funds to buy nodes, with notification by early-April, and a May-June order period. Equipment would be delivered in September, with the machine live by November. Chris asks the Committee for feedback as to any better timeline.

Raj Bose remarks that the development of this annual purchase cycle and timeline is a significant first step, and asks whether a faculty group will reconvene to review this timeline. Chris responds that it would be ideal if the University bought a pool of nodes every year and sold them back to researchers. However, although it is difficult to identify what unit would fund the initial purchase. Victoria Hamilton adds that typically SRCPAC has issued an RFP one year, followed the next year by an expansion. The Faculty RFP committee was quite useful, although it did consume time.

Ryan Abernathey recognizes that the majority of moneys originate from the schools, and so wondered if schools could buy extra nodes during the purchase rounds. Instead of offering funds through start-up packages, they could offer the nodes directly. Chris thinks that this will be tricky. Gaspare LoDuca suggests we save this for a discussion at the Research Computing Executive Committee, and Victoria agrees. Ryan notes that Lamont-Doherty Earth Observatory has done this on a much smaller scale, although this is clearly more difficult to do for all of Morningside and Manhattanville. Bob Mawhinney notes that there is a great deal of specificity for what kinds of nodes will be needed, which will present an additional challenge for anticipating need.

Jochen Weber asks what the Data Center's maximum capacity for power consumption, and how much larger we can grow. Rob Lane responds that this requires planning – we can complete a new expansion round, but then will need to integrate more cooling in order to accommodate a further expansion. Thus, it is difficult to anticipate what the subsequent year's capacity will be. An incremental 100 nodes would definitely be possible, though. Chris notes that the wall is closer than we would like, and there is not much room, although hopes that we can push this limit. Chris wants to preserve the possibility of an annual expansion. Rob adds that historically it has happened every 18 months.

Jochen additionally asks whether last year's prices will be fixed. Rob says he is currently negotiating with the vendor to achieve similar prices to last year, with an RFP process a fallback option if needed.

Overview of Training Services

Chris transitions to a discussion of some available courses for data science teaching and training, and suggests we publicize what we have done and what is available. Chris introduces Mark Newton, Director of the Center for Digital Research and Scholarship, who speaks to a range of training opportunities run by or with help from the Libraries, including the GSAS Digital Orientation, Open Labs, and Lynda.com. For more information regarding Mark's presentation, please view the corresponding handout (posted to the SRCPAC website).

Chris notes that this is an especially impressive portfolio, and asks whether there are any questions. Ryan Abernathey asks how these resources are being publicized. Mark responds that there is not yet a single venue for communicating these offerings, and Halayn Hescock notes that the compilation of these resources into this handout is a valuable first step. Rob Cartolano says that the Libraries is continuously interested in improving its outreach functions, and is very eager to receive feedback and suggestions. Chris notes that it would be advantageous to adhere to a standard schedule so that faculty can advise their students and postdocs on when to expect training opportunities. Ryan suggests that we further engage departmental curriculum committees as aids for publicizing these offerings to their students (especially incoming graduate students).

Raj Bose asks how the data management courses are going, as this is a pervasive challenge across the University, and also asks whether multiple disciplines are attending. Amy Nurnberger responds that there has been a regular schedule for these courses, and she has also been visiting departments that request her attendance. Workshop attendance is hindered by a lack of communications. Recently, the Senate suggested a number of email listservs aimed at promoting these services.

Kyle Mandli is surprised by the Business School's ample participation in these services, and asks what they are doing to publicize these services. Jochen Weber notes that the Business School has an urgent drive to utilize novel technologies, and this culture of awareness may have driven them to participate more fruitfully in CDRS' workshops. This culture should be promoted at all Columbia units as well. Barbara Rockenbach suggests that this is due to the existence of a dedicated Business Library with dedicated staff who have a visible presence to push these opportunities to students. Chris suggests that some units do publicize, while others don't know what they don't know.

Gaspare LoDuca echoes CUIT experience that communications are significantly difficult across Columbia. Emails do not work to make individuals aware of resources available to them Rob Cartolano adds that targeted communications are key.

Paul Blaer notes that the most activity is search-driven, and the Columbia website search tool is not yet fully adequate. It would be advantageous to optimize this search feature, such as by directly linking this to the CLIO service in order to advertise all resources.

Mark then introduces Ryan Abernathey, who spearheaded a large Software Carpentry two-day workshop on Python, MatLab, and R, which attracted 132 attendees. For more information on the breakdown of the 132 attendees, please see the corresponding presentation slide.

The aspiration for the SWC workshop emerged from experiences within the Department of Earth & Environmental Sciences, which does not expect its graduate students to know research computing skills, although these are becoming more and more important. Ryan has observed a substantial gap between students who do and do not have these skills, which constrains what research they can pursue and their career mobility. Ryan initially began teaching his own courses, exclusively in Python, although this was a substantial amount of work and was not part of his formal teaching. He became aware of SWC through his Open Source activities. SWC utilizes Open Source software, and is constantly updating and improving it. More information regarding SWC and the two-day workshop at Columbia can be found on the corresponding presentation slides.

Ryan explains we have the opportunity to become an Institutional Partner with SWC, which will cost $7.5k annually, which affords us the opportunity to train our own instructors and serve on the SWC advisory board. Ryan suggests that SRCPAC think about the utility of establishing this partnership.

Jochen Weber asks whether RCS would initiate the partnership so that their staff can be the trained instructors. Ryan suggests that the Libraries do this instead. Jochen clarifies that trained instructors should be administrators, not, say, postdocs who are transient by nature. Chris does not want departments to bear this new workload.

Ryan suggests a separate meeting with everyone involved in training – CUIT, Libraries, Data Science Institute, Computer Science – to develop a one-stop web portal to unify all training opportunities, to guide researchers towards the resource best for them, which can also be communicated to departmental curriculum committees. Chris suggests establishing a special Subcommittee towards this effort with clear objectives and deliverables. He will follow-up shortly regarding this Subcommittee.

Victoria then speaks to the Data Science Institute's five-day Data Science Bootcamp, which received 118 applicants for only 30 spaces. More information regarding this Bootcamp can be found on the corresponding presentation slides.

Chris notes that we would very much like to transition this responsibility over to the Libraries and ascertain what the Libraries are able to do.

Data & Society Task Force

Chris asks Victoria to review the history of the Data & Society Task Force. In 2015, President Bollinger asked that the Task Force be created to determine the strategy for the DSI. Currently, the Task Force is positioning data science as a focus of the upcoming capital fundraising campaign. Mary Boyce (SEAS Dean) and David Madigan (A&S EVP and Dean) chair the Task Force. The Task Force includes a Cyberinfrastructure subcommittee, led by Mike Purdy and Gaspare LoDuca, and a faculty advisory committee including Chris, Ryan Abernathey, and Louisa Gilbert. More information regarding this Task Force and the subcommittee's recommendations can be found on the corresponding presentation slides.

Publications Reporting

A compendium of publications emerging out of work hosted by Habanero and/or Yeti is very important in order to persuade research leadership to further support these efforts, as well as to demonstrate to the NIH the value of the G20 award for a Research Data Center. All SRCPAC members will soon receive an email asking them to report this information in advance of the Spring 2017 Research Computing Executive Committee. Publications can be reported at any time by emailing srcpac@columbia.edu.